International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 12 3808 – 3811

Optimum Cluster Labeling and Document Clustering for Forensic Analysis

Miss. Sushadevi Shamrao Adagale
Computer Engineering
Baramati, Pune, India,
susha0810@gamil.com

Mr. Vairagar Sagar G.
Computer Engineering
linePune, India
sagar.vairagar@gmail.com

Ms. Shubhangi Sagar Vairagar.
Computer Engineerin
Pune, India
vairagarss@gmail.com

Mr. Prof. Amrit Priyadarshi
Computer Engineering
Bhigawan,Pune, India
amritpriyadarshi@gmail.com

***Abstract*—**Document clustering or unsupervised document classification is an automated process of grouping documents with similar content. Document clustering is an important task in many Information Retrieval systems. Also document clustering Algorithms can help in discovery of new and useful knowledge or novel class from the documents under analysis. This knowledge or novel class is very important issue while handling forensic analysis. Digital Forensic Investigation is the branch of scientific forensic process for investigation of material found in digital devices related to computer crimes. In computer forensics, hundreds of thousands of files per computer are examined. Hence methods for automated data analysis, such as clustering are required.

Labeling large data sets with clusters bridges the effective cluster analysis to the large data set. Labeling irregular shaped clusters, distinguishing outliers and extending cluster boundary are the main problems in this stage. We address these problems and propose a cluster labeling algorithm which is very intuitive and easy to use

Keywords: *Computer Forensics, Clustering Cluster labeling,*

*****

I. INTRODUCTION

The latest edition of the annual Internet Trends report finds continued robust online growth. This survey shows that in near future digital information system may grows to 7-8 zettabytes of data up to year 2015. This survey itself shows that digital document handling is very important but complex task exist.

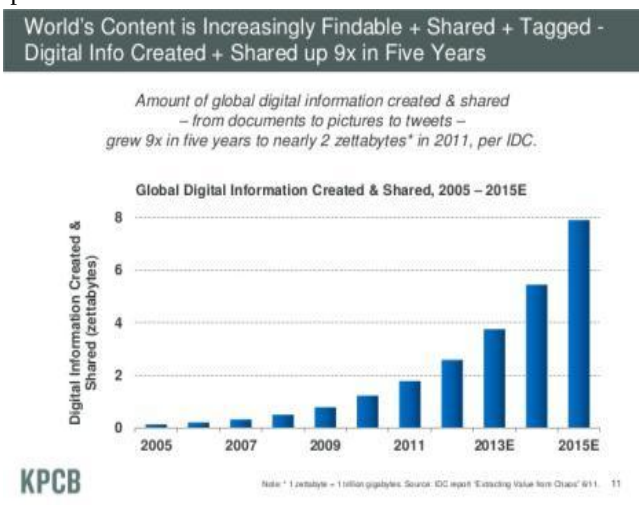

Figure 1.1 Digital Information System Survey

This large amount of data has a direct impact in Computer Forensics, which can be broadly defined as the discipline that combines elements of law and computer science to collect and analyze data from computer systems in a way that is admissible as evidence in a court of law. In our particular application domain, it usually involves examining hundreds of thousands of files per computer. This activity exceeds the expert's ability of analysis and interpretation of data. Therefore, methods for automated data analysis, like those widely used for machine learning and data mining, are of paramount importance. In particular, algorithms for pattern recognition from the information present in text documents are promising, as it will hopefully become evident later in the paper.

From a more technical viewpoint, our data sets consist of unlabeled objects—the classes or categories of documents that can be found are a priori unknown. Moreover, even assuming that labeled data sets could be available from previous analyzes, there is almost no hope that the same classes (possibly learned earlier by a classifier in a supervised learning setting) would be still valid for the upcoming data, obtained from other computers and associated to different investigation processes. More precisely, it is likely that the new data sample would come from a different population. In this context, the use of clustering algorithms, which are capable of finding latent patterns from text documents found in seized computers, can enhance the analysis performed by the expert examiner [1].

In this work, however, the clustering and labeling tasks are separated into two independent processes. First, a cluster partition of the data set is produced by a fully unsupervised

 ISSN: 2321-8169
3808 – 3811

clustering algorithm. Then, given a small set of labels (also referred to as prototype of labeled seed), a cost matrix is computed based on the distribution of labels throughout the clusters. The cluster labeling objective is then formulated as an assignment problem that is solved using the Hungarian algorithm [2]. Thereby, an optimum cluster labeling given the labeled seeds is ensured.

II. LITERATURE SURVEY

Sr No	Paper/Author	Advantage	Disadvantage
1	Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection. Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka 2013	Presented an approach that applies document clustering methods to forensic analysis of computers seized in police investigations.	It creates number of cluster. But to do analysis it not provides cluster labels which analysis easy.
2	A semi-supervised cluster-and-label approach for utterance classification Amparo Albalate1, Aparna Suchindranath1, David Suendermann2, Wolfgang Minker1	Semi-supervised cluster-and-label algorithm for utterance classification.	Labelling algorithmis very efficient but not work with other clustering algorithm like CSPA, Hirechical etc.
3	Clustering of documents in forensic analysis for improving computer inspection. K.Pallavi, S. NagarjunaReddy, Dr. S. Sai Satyanarayana Reddy	Clustering is often one of the first steps in data mining analysis. The partitioned K-means algorithm also achieved good results when properly initialized. Considering the approaches for estimating the number of clusters, the relative validity criterion known as silhouette has shown to simplified version.	We observed that in this clustering algorithms this only handles clusters formed by either relevant or irrelevant documents
4	Forensic Analysis of the Tor Browser Bundle on OS X, Linux, and Windows Runa A. Sandvik	Work on Operating system level. And updates as per training	Not handles clustering of documents
5	A comparative study on unsupervised feature selection methods for text clustering	Four unsupervised feature selection methods, DF, TC, TVQ, and a new proposed method TV are introduced.	Only works for text clustering and clustering labeling not applied.

TABLE NAME: LITERATURE SURVEY

III. ARCHITECTURE

It will also include the study of partitioned algorithms such as K-means & K-medioids; hierarchical algorithms – single, complete, average link and cluster ensemble based algorithms known as CSPA. The new proposed main work is to develop a novel hierarchal algorithm for document clustering in Computer Forensics, which provides labels for each clusters so that, it is very easy to do analysis of data for every cluster.

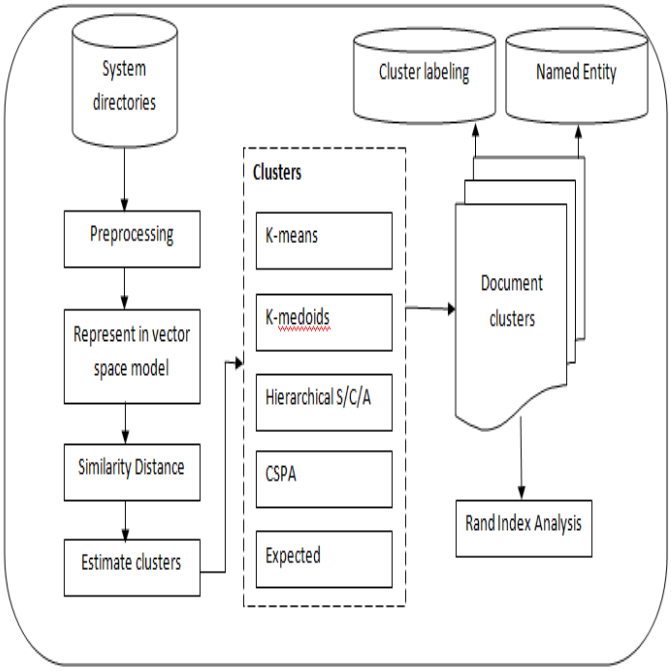

Figure 1.2 Architecture

Here we want to cluster various documents which include various steps:

1. Pre-processing – In this process we remove unwanted data or manipulate data which helps to cluster algorithm work efficiently. This process includes stop word removing, stemming, pruning etc.

2. Represent data in vector – Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers.

3. Similarity distance – This takes input as data vector and using similarity algorithm we get weight matrix for data vector or documents.

4. Estimate clusters –

1.K – means - K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

2.CSPA – Using METIS algorithm to partitions the data in similarity graph to obtain desired number of clusters.

3.K-medioids - The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoid shift algorithm.

4 .EM – An iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step.

5 Labeling –

Here we get input as various cluster and along with its members. We presented a semi-supervised cluster-and-label approach to classification of utterances has been presented. Then, the take output cluster partition to this algorithm as well as a small set of labeled prototypes (also referred to as labeled seeds) are used to determine the optimum cluster labeling related to the labeled seed. We formulated the cluster labeling problem as an assignment optimization problem whose solution is obtained by means of the Hungarian algorithm.

6 Project Analysis –

1. Label accuracy
2. Rand index

IV. DOCUMENTS CLUSTERING [1]

A. *Estimating number of clusters:*

In order to estimate the number of clusters, a widely used approach consists of getting a set of data partitions with different numbers of clusters and then selecting that particular partition that provides the best result according to a specific quality criterion. Note that, by choosing such a data partition, we are performing model selection and, as an intrinsic part of this process, we are also estimating the number of clusters. A widely used relative validity index is the so-called silhouette[6]

B. *Clustering Algorithms*

Given the training data, $X_T=X^{(t)}_T \cup X^{(u)}$, the set $Y^{(t)}_T$ of labels associated with the portion $X^{(t)}_T$ of the training set, the set K of labels for the k existing classes[1], and a cluster partition C of X_T into disjoint clusters, the optimum cluster labeling problem is to find a bijective mapping function

$$L: C \rightarrow K, K= \{1, 2, 3, \ldots, k\}$$

That assigns each cluster in C to a class label in K, while minimizing the total labeling cost. This cost is defined in terms of the labeled seed ($X^{(t)}_T$, $Y^{(t)}_T$) and the set of clusters C. Consider the following matrix of overlapping products N:

$$N=\{ \begin{matrix} n_{i1} & n_{i2}.....n_{ik} \\ . & . \quad . \\ n_{21} & n_{22}.....n_{2k} \\ . & . \quad . \\ . & . \quad . \\ n_{k1} & n_{k2}.....n_{2k} \end{matrix} \}$$

With constituents n_{ij}, denoting the number of labeled patterns from $X^{(t)}_T$ with class label y=I that fall into cluster C_j. The labeling objective is to minimize the global cost of the cluster labeling denoted by L:

$$\text{Total cost (L)} = \sum w_i.\text{Cost}(L(C_i)) \qquad (1)$$

Where $W= (w_1, , w_k)$ is a vector of weights for the different clusters. For example, it may be used if cluster sizes show significance differences among the clusters. In this paper, the weights are assumed to be equal for all clusters, so that $w_i = 1$,

The individual cost of labeling the cluster C_i, with class $L(C_i)$ is defined as the number of samples from class $L(C_i)$(in the labeled seed) that fall outside the cluster C_i i.e.:

$$\text{Cost (L } (C_i)) = \sum {}^{n}L(C_i)k. \qquad (2)$$

Applying Equation 2 to the total cost definition of Equation 1 yields:

Total Cost (L) = ∑ciєc ∑c!=cj ⁿL(Ci)k. (3)

In this paper, we applied the Hungarian algorithm to achieve the optimum cluster labeling in Equation 3. It requires the definition of the cost matrix $C_{[kxk]}$ whose rows denote the clusters and the columns refer to class labels in K. The elements C_{ij} denote the individual costs of assigning the cluster C_i to class label j, i.e. C_{ij} = Cost (L (C_i) = j). The reader is referred to [6] for further details about the assignment problem and the Hungarian algorithm.

V. CONCLUSION

We presented an approach that applies document clustering methods to forensic analysis of computers seized in police investigations. Also here we presented efficient clustering labeling algorithm which helps forensic analyst to handle cluster efficiently after getting the cluster labels. Here we showed how cluster labeling work and gives better labeling. By using these labels it's very easy to analyst to handle cluster.

REFERENCES

[1] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection – 2014,"

[2] Amparo Albalate1, Aparna Suchindranath1, David Suendermann2, Wolfgang Minker1, "A semi-supervised cluster-and-label approach for utterance classification,"

[3] K.Pallavi1, S.NagarjunaReddy2, Dr.S.Sai Satyanarayana Reddy , "CLUSTERING OF DOCUMENTS IN FORENSIC ANALYSIS FOR IMPROVING COMPUTER INSPECTION 2014 ,"

[4] Runa A. Sandvik runa@torproject.org , "Forensic Analysis of the Tor Browser Bundle on OS X, Linux, and Windows Tor Tech Report 2013-06-001," June 28, 2013

[5] Jianchu KANG, Jing YU, Zhongliang WANG, "A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering Luying LIU,"

[6] Tarek F. Gharib1,2, Mohammed M. Fouad3 , Abdulfattah Mashat1 ,Ibrahim Bidawi1, "Self Organizing Map -based Document Clustering Using WordNet Ontologies," 2012

[7] E. R. Hruschka, R. J. G. B. Campello, and L. N. de Castro, "Evolving clusters in gene-expression data," Inf. Sci., vol. 176, pp. 1898–1927, 2006.

[8] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, andM. S. Oliver, "Exploring forensic data with self-organizing maps," in Proc. IFIP Int. Conf. Digital Forensics, 2005, pp. 113–123.

[9] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results," Digital Investigation, Elsevier, vol. 4, no. 1, pp. 49–54, 2007.

[10] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework," Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137, 2009.